Getting Started

Table of contents

1 Installation....2
2 Plug-ins.... 3
3 Usage....4
4 Directory Structure 4
5 Ingest Logging Format....4

1. Installation

```
1. Copy oai-resource to a location of your choosing.

2. Configure the env.properties file.
This file should be stored in a global location, as all oai resource
harvesting projects will most likely share the same file.

The following configurations are available:

* oairesource.tapePrefix=tape
- This is the prefix assigned to the generated xml tape. (REQUIRED)

* oairesource.gzip=false
- Defines whether or not gzip compression should be used for generated
archives. Not currently supported.  (REQUIRED)

* oairesource.maxsize=100000000
- Defines the maximum size of each arc file. Default is 1GB. (REQUIRED)

* local.datastream.prefix=info:lanl-repo/ds/
- Defines the datasream prefix to assigned to the UUID associated
with each resource.  This should be a local specific uri prefix. (REQUIRED)

* local.openurl-referrer.id=info:sid/library.lanl.gov
- Defines the service id used for external reference. (REQUIRED)

3. For each oai resource harvesting project, a project properties
file needs to be defined.  An example of this file can be found
in etc/config/project.properties.  Typically, project property
files will be stored at the root of the project directory.

The following configurations are available:
* oairesource.projectName=EPRINT
- Project name, used as a prefix for arc file names (REQUIRED)

* oairesource.baseurl=http://libeprints.open.ac.uk/perl/oai2
- The OAI BaseURL to the OAI PMH Repository you wish to harvest
resources from. (REQUIRED)

* oairesource.from=
- OAI From Date Variable, overridden by last harvest file property,
if present (optional)

* oairesource.until=
- OAI Until Date Variable (optional)

* oairesource.sets=
- OAI Set Variable (optional)

* oairesource.metadataPrefix=oai_dc
```

```
- OAI Metadata Format of the data to be queried (REQUIRED)

*
oairesource.ingestPlugin=gov.lanl.ingest.oaitape.simple.DCIdentifierProcessor
- OAIResource Plug-in to be used to process the OAI Result set.
See plug-in example section below. (REQUIRED)

* oairesource.projectDir=~/dev/tmp/eprint/
- The data directory for project resources to be written.  The xmltape
file, harvested resource arc files, and log files will be written to
this directory.
```

2. Plug-ins

```
With this release, a few examples of the plug-in interface are provided:

Package: gov.lanl.ingest.oaitape.simple
* DCFormatProcessor
In this example, when provided oai_dc metadata, this processor obtains
the value of the dc:format element. If the resource is a pdf, tif, jpg,
or mp3, the resource will be harvested and written to the arc file.

* DCIdentifierProcessor
In this example, the URI defined in dc:identifier references a
human-readable
splash page describing a resource. The URI doesn't directly reference the
resource itself.  The processor then parses the splash page for links to
pdf, tif,
jpg, or mp3 files (these would be considered the resources) and downloads
them.

* DCIdentifierPDFProcessor
Similar to DCIdentifierProcessor, but only harvests PDF resources.

NOTE: Resource harvesting and processing will typically be
repository-dependent.
You'll want to create implementations which support the resource structure
of
your repository.  The provided implemenations are intended for
demonstration
purposes only.

Package: gov.lanl.ingest.oaitape.aps
* DidlSigProcessor
A plugin for DIDL metadata with xmlsignatures.

Package: gov.lanl.ingest.oaitape.modoai
* DidlNoSigProcessor
A plugin for DIDL metadata with/out xmlsignatures.
```

3. Usage

```
1. From the bin directory, run the following command:

Syntax:
sh ./OAIResource.sh <absolute path to env.properties file>
<absolute path to project.properties file>

Example:
sh ./OAIResource.sh /Volumes/UserData/admin/OAIResource/etc/env.properties
 /Volumes/UserData/admin/OAIResource/etc/aps/aps.prorties
```

4. Directory Structure

```
bin
 |- env.sh - System environment set-up, called by main script
 |- log4j.properties - Defines debug level for application
 |- OAIResource.sh - Main OAIResource Application
etc
 |-config
 |  |- env.properties - Defines global variables
 |  |- project.properties - Defines project variables
 |-examples - Samples for project/env set-up
lib - Necessary libraries for application
```

5. Ingest Logging Format

```
A record is written to ok.csv for each resource which is successfully
harvested to an arc file.

Log File Format:
tape_record_id, arc_id, arc_date, ref, derefXPath, sourceURI,
digest, localIdentifier

Example:
oai:aps.org:PhysRevC.71.065801,APS_fd412758-85e0-4ac0-a104-b45a9fa82dae,
20050801172118,http://oai.test.ridge.aps.org/filefetch?identifier=PhysRev
C.71.065801&component=metadata&description=apsmeta&format=xml,//didl:
Component[0]/didl:Resource[0]/@ref,,urn:sha1:GcmDeJ7jqqZENRdO9EGAbZ4D9aY=,
info:lanl-repo/ds/6040d77b-15b3-42b4-9bbc-e0332cd6f3fa

Column Descriptions:
* tape_record_id - OAI Record ID
```

```
* arc_id - Name of the arc file in which the resource resides
* arc_date - Date on which the resource was written to the associated arc
file
* ref - URL reference from which the resource was harvested (XPath using
ref)
* derefXPath - XPath from which the ref value was originally obtained
* digest - Generated digest of harvested resource
* SourceURI - URL reference from which the resource was harvested
* localIdentifier - Locally generated UUID
```